EPIQ IQC, PCE-I-00-96-00002-00, Task Order No.839

Madagascar Biodiversity Analysis Platform Development: Report on Module 3

Prepared by
Claire Kremen

with contributions from
David Lees
John Fay
Lanto Andriamampianina

International Resources Group, Ltd.
1211 Connecticut Avenue, NW, Suite 700
Washington, DC 20036 USA

Prepared for:
USAID/Madagascar

July 6, 2001

Environment Policy and Institutional Strengthening Indefinite Quantity Contact (EPIQ)

Partners: International Resources Group and Winrock International

Subcontractors: PADCO, Management Systems International, and Development Alternatives, Inc.

Collaborating Institutions: Center for Naval Analysis Corporation, Conservation International, KNB Engineering and Applied Sciences, Inc., Keller-Bliesner Engineering, Resource Management International, Inc., Tellus Institute, Urban Institute, and World Resources Institute

Table of Contents

Overall Goal

The Plateforme d'Analyse de la Biodiversité de Madagascar (PDA) Project aims to establish a spatially-explicit biodiversity database in Madagascar that can be maintained and updated in Madagascar by Malagasy technicians and can be used to conduct a variety of conservation planning analyses. This database will add value both to the existing biodiversity information available in Madagascar and to the national capacity in Environmental and Geographic Information Systems.

The overall goal is to create a tool that improves environmental decision-making by assembling hitherto scattered biodiversity data and enabling users to conduct advanced spatial analyses for conservation planning and environmental management.

Project rationale and justification

Over the past ten to fifteen years, the scientific community has made substantial progress in cataloguing Madagascar's biodiversity. However many gaps remain in our knowledge of species distributions across Madagascar. In addition, the existing information is not centralized in any one database. Consequently, it is difficult to use these data efficiently for conservation planning and management. In 1995, a critical workshop was held that brought together key experts in biodiversity and socio-economic studies to identify the priorities for conservation and further research. While this workshop advanced the conservation planning process for Madagascar, the lack of consistent, reliable, and comprehensive databases on species distributions continues to constrain progress in determining on-the-ground priorities for conservation work. There is a tremendous need for a national, spatially-explicit, biodiversity database that can capitalize on the vast store of biodiversity information that has been generated recently and use this information for conservation analyses and priority-setting. WCS/CCB has developed a prototype database and mapping program for Madagascar that could serve as the cornerstone for the development of a national Malagasy program.

Report on activities covered under Module 3

(for other modules see respective reports)

Goal 1: Translating PDA tools into Visual Basic

We have streamlined the interface of the ArcView 3.2 PDA extension and added a few more features to facilitate range processing and validation. We have also developed a stand-alone PDA tool using Visual Basic so that range analyses can be done independently of ArcView. The current version of the Visual Basic program contains most of the features that the ArcView version and also has limited GIS capabilities.

Goal 2: Assessing data holdings and participation by other institutions

We visited or contacted the following institutions and individuals with the following results.

Institution	Individual	Interest in participating	Taxonomic group	Digital database available	Requested funds?	Number of species that are "ready"	Other notes
AMNH/ WCS	Stiassny/ Loisselle	enthusiastic	freshwater fish	yes	yes	~45	Four fish families[1]
AMNH	Raxworthy	possibly in the future	amphibians /reptiles	yes	yes	?	Funds requested not appropriate
U Michigan	Nussbaum	did not respond	amphibians /reptiles				
MBG	Lowry/ Schatz/ Birkinshaw	enthusiastic	Plants	yes	no	~675[2]	Tropicos database
Kew	Hoffman et al.	not yet known	Plants	yes	no	~690[3]	
Harvard	Alpert	enthusiastic	Ants	yes	no	?	Filemaker database of 20000 records
CAS	Fisher	yes	Ants	yes		?	Biota database
Keene College	Zjhra	enthusiastic	plants: Bignoniaceae	no	no	70	
FMNH	Goodman	enthusiastic	small mammals	yes	no		95% of taxa are ready; the rest can be done soon
	Wilme/Good-man	yes once data is published	Birds	yes	no		100,000 records
ZICOMA	Vony Ramino-arisoa	enthusiastic	Birds	yes	no	?	computer that contains the database is down currently
WWF-Madagascar	Achille Raselimana-na	yes	various taxa	no	no	?	are currently developing database
ICTE/MICET	Liva Rajoharison	yes	Various taxa	no	yes	?	Are currently developing database
TPF	Aristide Andrianari-misa/Zarasoa	yes	Birds of prey	no	no	?	Are currently developing database

1. Cichlidae, Apochidae, Anchariidae, Ariidae
2. Seven endemic plant families ("complete"), Araliaceae, papilionoid legumes, and other plant families (all other familiess partial)
3. Palms ("complete"), Papilionoideae, Caesalpinoideae, Mimusoideae (all other families partial)

Goal 3: Conversion module from NOE database format

In addition to dBase, text, and Excel formats, the PDA tool will also be adapted to read databases exported by NOE once the NOE format is made available. To date, the required data has not been made available by ONE, because they apparently have not been able to export it in an ASCII-file format, although the NOE manual states that this option (standard on most software) is available. If they could export it in an ASCII-file format, it would be easy to bring it into the PDA format.

Goal 4: Adding additional data (butterflies)

During Module 3, a large effort by DCL was put into primary data capture from MNHN, Paris and BMNH, London, focusing entirely on butterflies. Over 4000 new records were compiled prioritising species (about 111) for which representation in the prior version of the PDA database was weak. For about 95 of these species, museum data is now reasonably complete. In contrast to the previous work, emphasis was placed here on specimen-level data capture (including unique numbering of specimens, full label capture and the highest quality taxonomic re-evaluation, all of which necessitate resorting of drawers). This level of databasing is justifiable because it results in data that is of lasting durability for PDA. Taxonomic coverage included near completion of Hesperiidae for both museums, and a considerable number of Acraeinae from both, some Pieridae from Paris, all Riodininae from London, some Papilionidae from both museums, and some other species scattered throughout the Nymphalidae in the genera *Charaxes*, *Neptis*, *Cymothoe* and *Amauris*. Specimen-level databasing will eventually be required for all species for which PDA already has partial label data, particularly Satyrinae (100 spp.) in London and Paris, Pieridae (30 spp.) and Lycaenidae (44 spp.) in London, Acraeinae (14 spp.) in Paris, and certain pierid genera such as *Belenois*, *Leptosia* and *Nepheronia* for which museum data at any level is mostly lacking. A number of other museums with smaller but important Madagascar collections (Oxford, Senckenburg, Saarbrucken) should also be visited in the future to capture data. Some field specimen data was also entered, mainly for *Acraea*. All records were georeferenced.

Goal 5: Checklists (butterflies and mammals)

Synonymic taxonomic checklists in electronic format for described Madagascan butterflies (300 species) and mammals (142 species; 165 subspecies) were prepared by DCL. The most important fields in this database were genus, species, subspecies, taxonomic authors and dates, type locality, type information including deposition in which museum, and type reference and principal synonyms. Mammals have been well typified in the literature and so most of the type information could be obtained; butterflies, in contrast, have not, so most of the type information is preliminary. The checklists are up to date for all described species to early 2001 including several newly described taxa. The revised butterfly checklist has now been submitted as a chapter of a book on Madagascar; therefore much of the work involved in its electronic compilation (several weeks, in fact) was included *gratis* to the project. It is emphasized that electronic checklists are not final but dynamic products, that will need continual updating and maintenance. They are key to accurate distributional data and any applied use of the data.

Other Activities

Three additional activities that were not described in the Module 3 proposal were carried out because they constituted opportunities that could advance the project.

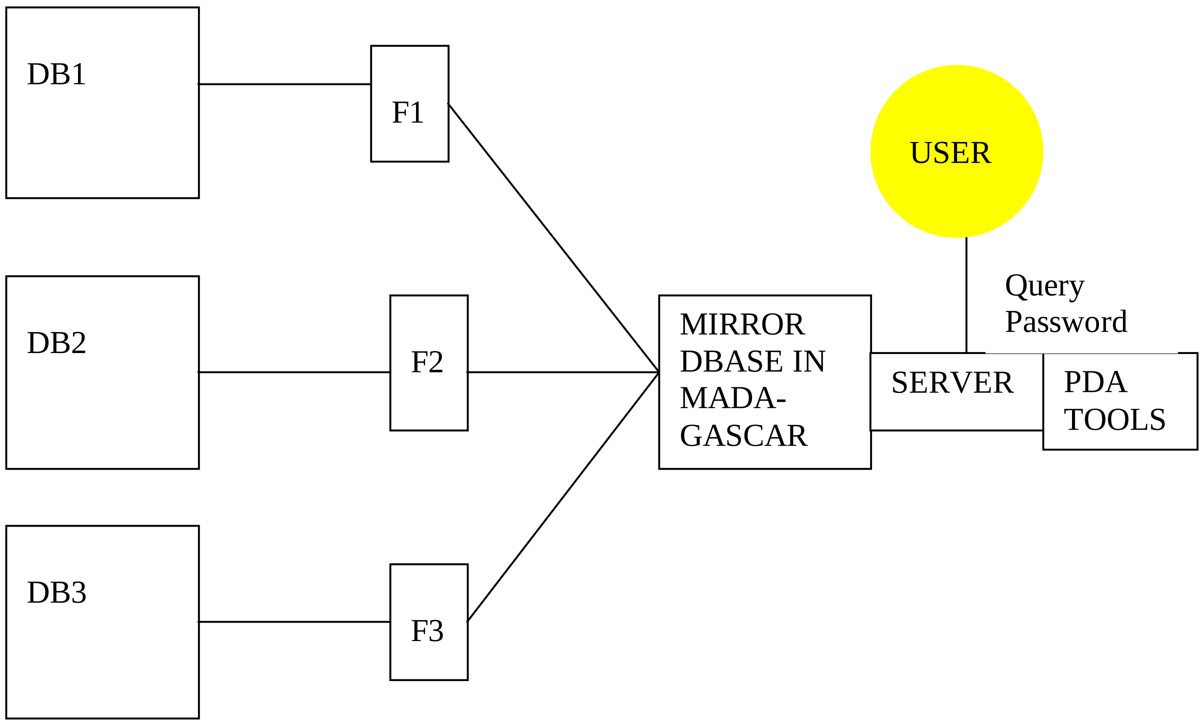
A visit by CK to CABS at CI-Washington proved highly productive. A demo was given of the existing PDA tools. Discussions were held with Silvio Olivieri, Gustavo da Fonseca, Lee Hannah, Carlos Galindo-Leal, and Crispen Wilson (BCIS). CI-CABS wishes to develop a web portal for Madagascar that would allow access to environmental and socio-economic information. They are eager to work with others in developing the portal; essentially they wish to catalyze/facilitate the development of it. They also have substantial expertise in their GIS laboratory and are willing to help in dealing with technical issues. Silvio wishes to work with Paul Williams at the Natural History Museum, for example, to recreate the Worldmap conservation planning tools in a more accessible format.

We worked with the Primate and Other Mammals Groups at CAMP to use the PDA to estimate extent of occurrence (e.g. area of polygon around point distributions) and area of occupancy (e.g. area of range prediction using latest available forest map, e.g. 1985, since IEFN data not yet available) for these species. The tool was successfully used to calculate areas of occurrence and occupation for many lemur and carnivore species. As for small mammals, it appears that for several species, the distribution ranges produced by the PDA did not match the real distribution and were not representative of the geographic truth. This was because several recent publications and unpublished reports were not yet incorporated into the PDA database. So although the ranges match for some species, the mammal specialists at CAMP decided not to use the tools in order to keep the mode of calculation uniform for all taxa. However, in general, all the people that were introduced to the PDA during the workshop seem very excited and find the tool very useful.

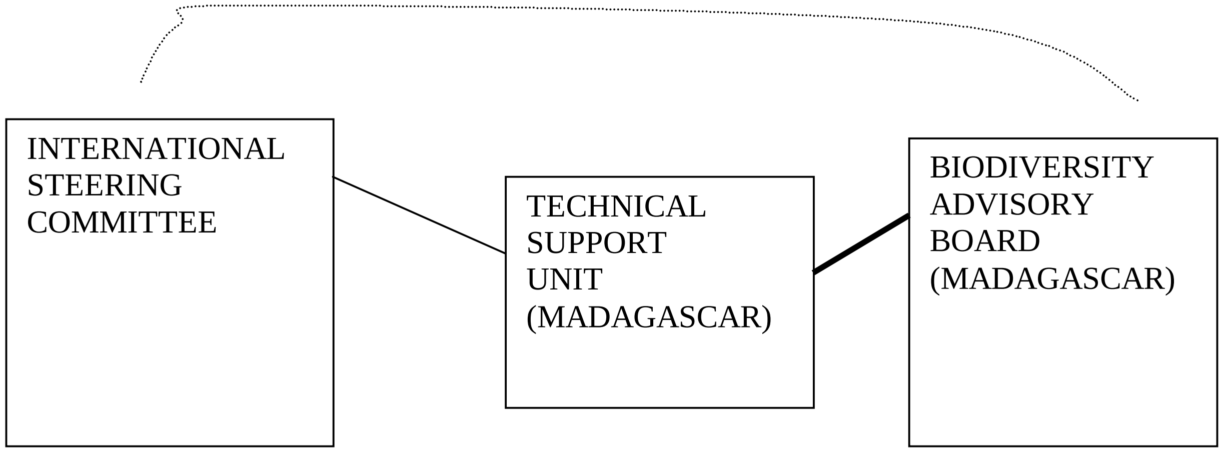
On CK's visit to Madagascar (April 7 – 21), financed by WCS and therefore *gratis* to the IRG-funded project, a new plan was developed for the technical and institutional structures needed for PDA (see diagrams below and Annex 1). This led to the development of the last funding proposal from PAGE (Module 4).

Attachment 1. Technical and Institutional Outline Concepts for PDA.

Technical Plan: A user with internet access anywhere in the world can visit the PDA website, and access biodiversity and other data layers for Madagascar via a password that determines the user's access level to information. The query passes in real time via the server to all contributing databases to collect available data responding to the query. The data from each institution (DB1, DB2…) passes through a filter (F1, F2…) which allows data from disparate sources to be collated into a single format for use in the PDA toolset, which is available to the user at the website. This same data also exists as a "mirror database" located on a server in Madagascar. The mirror database provides a backup copy of the dataset and allows access to the data even when other servers are down. It can be updated automatically on a regular schedule. Institutions control the level of access to their data through an access code, according to guidelines established by the International Steering Committee.



Institutional Plan:



INTERNATIONAL STEERING COMMITTEE

TECHNICAL SUPPORT UNIT (MADAGASCAR)

BIODIVERSITY ADVISORY BOARD (MADAGASCAR)

MEMBERSHIP:

DATA HOLDERS ; MALAGASY MINISTRY REPRESENTATIVES; BAB Representatives	DIRECTOR GIS EXPERTS SOFTWARE ENGINEERS BIODIVERSITY EXPERTS	NGOs & Gos IN MADAGASCAR LOCAL REPRESENTATIVES OF DATA HOLDERS

FUNCTIONS:

DEVELOP GUIDELINES FOR DATA ACCESS, DATA SHARING, DATA VALIDATION, TAXONOMIC AUTHORITY, GAZETEER, & SUST-AINABLE FINANCING FOR PDA	PROVIDE TECHNICAL SUPPORT TO NGOs & Gos ON DATA ANALYSES FOR CONSERVATION PLANNING, INCLUDING TRAINING GIS STAFF OF USER ORGANIZATIONS; PROVIDING A HELP SERVICE; PROVIDING DATA & TOOLS TO NON-INTERNET USERS	PARTICIPATE IN DEVELOPMENT OF INTERNATIONAL GUIDELINES WITH ISC; DEVELOP GUIDE - LINES FOR DAILY MANAGEMENT OF TSU WORKLOAD; IN ASSOCIATION WITH ARSIE, HANDLE ISSUES RELATED TO DATA ACCESS WITHIN MADAGASCAR

Attachment 2. Report on Module 3 for PAGE: Butterfly/mammal consultancy

Prepared by David Lees
Jan-Jul 2001 (51 days)

The TOR of this consultancy were to:

1) Capture museum butterfly data into PDA from the museums:

    BMNH

    MNHN

2) Incorporate butterfly field data databases into PDA and correct existing geographic info where inaccurate

3) Prepare complete taxonomic check-lists for

    All described Malagasy butterflies

    All described Malagasy mammals

4) Georeference the data and estimate geo-errors using principal digital gazeteers

5) Demo Platforme d'Analyse and carry out questionnaire at botanical institutes:

    MNHN Paris

    RBG Kew

Additionally, enhancement was carried out to the mammal database in the context of Module 2.

(1). Adding additional museum data (butterflies) – 29 days

During Module 3, a large effort by DCL was put into primary data capture from BMNH, London and MNHN, Paris, focussing entirely on butterflies. It was decided to focus on butterflies because of the author's taxonomic expertise on this group and the possibility to obtain a comprehensive coverage distribution points in relation to available collections visited in London and Paris.

Over 4,000 records were compiled, prioritising species such as skippers for which geographic representation in the prior PDA database compiled in 2000 was weak (although databased, some of these records were excluded from the resulting file PDA butterflies.dbf because their historical data is too imprecise to be useful in PDA e.g. "Madagasacar"). For about 95 of the species comprehensively input this time, museum data can be considered reasonably complete across both museums. In contrast to the previous work, emphasis was placed during this consultancy on specimen-level data capture. This requires unique numbering of specimens, full label capture and high quality taxonomic re-evaluation, all of which necessitate geographic resorting of drawers. For the other approximately 200 species of described Madagascan butterfly, there is already a good geographic spread from new and previous input including field data (especially for the 84 described species of Satyrinae), and although label information was captured, specimen level numbering is not yet implemented. The butterfly database thus now represents a solid, reasonably comprehensive coverage nationally of described butterfly species consisting of over 12,000 validated records. The taxonomy has also been resolved to the highest standards presently available, in accordance with Lees, Kremen and Raharitsimba (2002, in press).

Taxonomic coverage

Whilst specimen-level databasing this does not necessary result in a greater coverage of geographic points, the data is of lasting value because no extra label input is required and taxonomy of any record can later be updated. Taxonomic coverage included near completion of Hesperiidae for both museums, a considerable proportion of Acraeinae from both, some Pieridae such as all *Colotis* from Paris, all Riodininae from London, some Papilionidae from both museums, and some other species scattered throughout the Nymphalidae particularly in the genera *Charaxes*, *Neptis*, *Cymothoe* and *Amauris*.

(2). Adding additional field data (butterflies) – 1 day

Over 1000 records (1052 representing described species) from Rhopaloc2.xls were completely integrated and georeferenced into PDA. These included recent field data from Masoala, Analamera, Anjanaharibe Sud, Ranomafana, Anjozorobe, Andringitra, and Andohahela (note that there are about 195 potential duplicates included from CK's database spplocal2BHM.xls previously input in Module 2 that may overlap with some previous records).

Over 3 days were also spent validating and otherwise cleaning the butterfly data

(3). Checklists (Madagascar butterflies and mammals) – 7 days

Synonymic taxonomic checklists in electronic format for described Madagascan butterflies (Papilionoidea and Hesperioidea, all 300 species: Mad Butt checklist.dbf) and mammals (all 142 species; 165 subspecies: Mad Mamm checklist.dbf) have been prepared. Explanations of fields included in these databases are included in the file "Taxonomic checklist metadata.xls" (Table 1).

Butterflies

The revised butterfly checklist has now been submitted as a chapter of a book on Madagascar (Lees *et al.* 2002, in press). The present consultancy has been highly synergistic with this piece of work in the sense that much of the effort involved in its compilation (several weeks) could be included *gratis* to the project whilst the consultancy time itself was spent on actual databasing.

Mammals

In contrast to the situation for butterflies, where most of the type information is still preliminary (Madagascar butterflies are in early stages of comprehensive typification), mammals have been well typified in the literature, and so most of the essential type information could be obtained. However, both checklists are up to date for all described species to early 2001 including several newly described taxa. At present, the literature references are included in appropriate fields in abbreviated fashion. Bibliographic sources for mammal and butterfly descriptions for the currently recognised species amongst described taxa have however been included in full. Relevant data from the mammal checklist (including principal synonyms with their type localities if known) has also been georeferenced and integrated into the mammal distributional database, creating 425 new records, 182 of these with coordinates.

Taxonomic link fields

At present a Linnean trinomial (the field Full_name1) is used as a link field to the distributional database. A more sophisticated system may ultimately be needed to deal with "uncleaned" data that includes synonyms, misspellings, etc. At present it is necessary to clean imported data, using the checklist as a lookup to get this link field correct in any new database that is being imported to PDA, before all the rest of the taxonomy can be joined automatically. Also if the trinomial is not known but is applicable, a binomial (Full_name2) can be used as the link field.

Standards

Although there are doubtless many ongoing improvements desirable to these two checklists, the bar has been set rather high in terms of the level detail incorporated. For example, vernacular names have been included (important for a wider audience). Endemism at genus, subgenus, species and subspecies level and for different broad geographic levels has been summarised. Detailed type and synonymic information has been included, etc. This should serve as a model on which to design checklists for other popular or less popular groups of organisms.

It is emphasised that such electronic taxonomic checklists are not final but dynamic products, that will need regular updating according to the latest literature, although they already include many elements such as references for description and type locality that are likely to be complete.

(4). Georeferencing [and chrono-referencing] – 5 days
As before, all new distributional records were georeferenced during this consultancy so that they could be mapped. Decimal latitude and longitude are the core descriptive fields that form the basis for linking a large amount of other metadata. A single electronic digital gazetteer, in Excel format, largely consisting of the US Board of Names (USBN) and MBG/TAN/FOFIFA gazeteers (digital sources are available on the web as the GEOnet Names server http://www.nima.mil/gns/html and on http://www.mobot.org/mobot/research/madagascar/gazetteer), was produced. This was intersected with the 30 arc second Digital Elevation Map for Madagascar and the population shapefile available from the CI 1995 workshop CD-ROM, for this purpose. Regularly used names were highlighted in bold in this file, because of the large problem with placename homonymy in Madagascar, to facilitate future use. This gazetteer still cannot be used indiscriminately but each record must be evaluated against other information. This product is called PDA_GAZ.xls. See File menu properties. Full metadata is included in a separate worksheet.

There was not the scope within this consultancy to implement the principal georeferencing methods presently used within an Arcview programming framework, such as tools to calculate implied errors in degrees minutes and seconds (DMS), distance and bearings and associated errors, nearest localities to a specified locality within a gazetteer, etc. However, the most important basic tools with inbuilt Help comments are implemented at the end of the gazetteer and in a separate worksheet. Further implementation should be considered for future work (see Recommendations).

How geographic errors were calculated (see previous consultancy report for more technical details)
Geographic errors are generally crude, especially for historical data. Here we use the concept of a **point and error radius** (in km.), rather than a **bounding box** as sometimes used for data that is expressed in lines of latitude and longitude. The reason for preferring the former are (1) it is simpler, requiring only 3 fields and (2) specimens are actually collected at a point whose coordinates are the parameters that need estimating.

If DMS or decimal latitude and longitude is already available in the data, an implied precision can be estimated (e.g. data may come in minutes, and this will translate into an about 1 km. radius depending on latitude). Most Madagascar GPS data is already latitude-longitude and does not needed datum conversion and, post April-2000, comes with an about 15 m. error (2 D navigation) or 7 m. error (3 D navigation), and order of magnitude improvement. If coordinates are not available, coordinates and their errors are estimated from the geographic name or other useful data that may be available, that is parsed into the Lcty_name field. If a distance and bearing from a known point are referred to, the approximate new coordinates and geographic error can be calculated from this. Often this new error circle can serve as a first approximation to a more refined GIS estimate of the locality.

If a particular geographic feature with known boundaries is referred to (e.g. country, province, lake, forest) and this information is already digitised in Arcview, coordinates and their error can be approximated. This can be done either by computing the centroid of the shape and its average radius as the square root of (Area/2*pi), or the mid-point and half-length for a linear feature. If a shape is available, this can be and is referred to in the metadata (such as the field Coord_Src), thus delineating the potential area of collection.

For older data such as exists within museums, all that may be available is a placename and often the collector and/or date. Often the locality can potentially be confirmed and fine-tuned by biographical or bibliographical research, especially if a map is available, or by reference to other data from the same collector. Often though there is not sufficient time to carry out such detailed historical research at input stage. Thus a first estimate of coordinates and error must be based on present day population placename from a digital gazetteer (with the caveat that villages tend to migrate over historical time). To this is added an estimate of how far the collector might have strayed from the named locality to capture say 75-95% of possible records. These “straying” distances have been here estimated in the first instance as 15 km (for pre-1939 data) and 5 km (for post-1939 data, assuming 15 km. to be an unacceptable precision for most modern data). Coastal data was not given a smaller error since the distance wandered may be the same (the previous PDA database was corrected in this respect). Errors can sometimes be refined down to about 1-2 km. using other information and GIS methods. Very vague locations such as tribal territories (not used for modern biodiversity data) were given a 25 km. error or, if larger, excluded from the PDA database. These figures represent an arbitrary series reflecting an increasing historical emphasis on precision, and are certainly subject to future re-evaluation. It may later become apparent that particular collectors’ data are unreliable, so that such data may need to be removed entirely from the database or coded as low confidence and filtered during the mapping process.

As elevation is essentially a single dimension, the error can simply be expressed as (+/-). This error (Elev_err) can be calculated, if such information is available, as Max_elev – Elev (where Elev is the mean elevation). If Elevation is derived from a digital elevation map, this is indicated in the field Elev_Calc by the letters "GIS".

How geographic errors can actually be used

At present PDA does not directly implement display of geographic errors, although this is desirable in future builds. These errors will be of direct use at such stage as probabilistic modelling is implemented to estimate probabilities of occurrence within gridcells. However, it is possible to display geographic errors from the Arcview interface in two ways.
1. Add the database as an event theme and click on the theme to display the legend editor. Select graduated symbol, and use Geog_err as then classification field. Set the size range to approximate to the errors involved. [This method is somewhat crude].
2. It is also possible to create a map of the error circles directly. Add the point theme as an event theme to the view, then use the Theme menu -> Create Buffers -> Select the theme -> Create buffers at a distance from the attribute field "Geog_err" to produce a new shapefile with radii of the appropriate sizes.

Time of collection and temporal errors

Just like the fields for coordinates, collectors, binomens or trinomens, unique codes associated with specimens and literature references, time of collection may be treated as a core descriptive field for biodiversity data. Such fields potentially link to other sets of information (respectively gazetteers, biographies, taxonomy, catalogues, specimen registers and field books, and bibliography): indeed this approach has been used in the field header metadata descriptions. Time is a single extra dimension but it is convenient to divide it into at least four independent fields. Where the span of time is known for a record, this can be described by a start and end date, and this is the system we presently use in PDA, adding an extra two fields. This is simpler and more directly represents the data than estimating a mid-time and error from such information. The start date may however be the only date available, in which case no attempt has been made to estimate its error, since that error is not of direct use in mapping. However, that field has been used to gather together Day, Month and Year into a single expression, where known. Date of accession to a museum may in some cases be the only indication of how recent a collection is, but that information is included in a different field (Mus_Acc_Da).

(5). Demos of PDA and questionnaires to botanical institutes – 5 days
This work is fully covered within a separate report (DL trip report.doc).

(6). Verification and enhancement of mammal database

The original mammal database was refined for the CAMP meeting in mid May 2001, as an additional service to Module 2. All previous records that included specimen level data (e.g. referred specimen numbers in the *Fieldiana* series) were parsed out to individual per-line records, in case of future needs as regards taxonomic changes. Secondary data source records were removed. During the validation procedure developed as an addition to the PDA extension, geographically incorrect records were identified and re-georeferenced, and unreliable records were removed. Volant mammals (bats) were removed from the primary database before this meeting as insufficiently comprehensive and difficult to map predictively with the existing tools and themes. Type localities for each species or subspecies from the mammal checklist were also georeferenced where possible and added to the distributional database. Records (such as of *Microcebus*) that are superceded by current taxonomy were reclassified as “sensu lato”. Ultimately, they need to be reassigned where possible to the new taxa.

The mammal data has now been divided into three separate databases (Table 1): validated non-volant mammals, volant mammals (not used in CAMP) and other records (including field guide data and secondary compilations, fossil data, morphospecies or taxonomically and geographically less reliable data). It is possible that for some purposes, these databases may need to be combined (they have the same field structure).

Recommendations arising from this consultancy

1. *Prioritise specimen-level or vouchered data*. The only durable biodiversity data (for those organisms that can readily be vouchered) in databases for multi-purpose use at a national scale is specimen-level, individually coded data. For organisms such as birds and lemurs, authoratively validated records should be prioritised, especially those supported by photographs, sound recordings, or genetic tissue samples. Although all literature records are valuable to variable degree as a scientific resource in a database, it is clear from only a year’s existence of the prototype mammal database that many bibligraphic datasources will become rapidly redundant and discarded. This is due not only to advances in georeferencing technology, but rapid advances in taxonomy. It is also now recommended that records from Nicoll and Langrand (1989) are removed, after consultation with Martin Nicoll who is unsure even of the validity of lemur records from individual reserves.
2. *Acquire up-to-date GIS themes and make them available*. New, much more accurate, georeferenced, themes are needed to maximise the value of taxonomic range predictions in GIS extensions such as PDA. Particularly important to acquire is a classified polygon vegetation coverage and a high resolution national DEM (such as the 90 m. one). More specific climatic

datasets than the Cornet map available on the MBG website will also be useful.

3. *Integrate georeferencing methods within PDA*. Future versions of PDA should have the potential to map errors and filter out records with unacceptably large errors. It should also be useful to have some georeferencing tools programmed in.
4. *Create a databasing front end suitable for local capacity building*. To date, data has quite simply been entered into Microsoft Excel to date by keyboarding or scanning routes and later converted to database format. It is recommended that to avoid potential garbling of existing data during data entry as well as to reduce file sizes, a database front end programmed in a widely available format such as Microsoft Access be used to facilitate data entry by trained Malagasy technicians. This could also have a number of geographic calculations incorporated as modules and lookup tables such as gazetteers and taxonomic checklists. Such a front end will become increasingly important as PDA moves towards a distributed server system. It is also likely that different datasheet form and relational designs and will be needed for different datatypes (e.g. terrestrial versus aquatic data), and for different ends (e.g. literature versus specimen data).
5. *Expand the butterfly database*. Specimen-level databasing will eventually be required for all species for which PDA already has partial label data, particularly Satyrinae (100 spp.) in London and Paris, Pieridae (30 spp.) and Lycaenidae (44 spp.) in London, Acraeinae (14 spp.) in Paris, and a few pierid genera such as *Belenois*, *Leptosia* and *Nepheronia* for which museum data at any level is still largely lacking. A number of other museums with smaller but important Madagascar collections (PBZT, Oxford, Senckenburg, Saarbrucken) should also be visited in the future to capture data. There is also limited field data to date from Western and Southern Madagascar.
6. *Expand the mammal database*. A more comprehensive input of appropriately prioritised recent literature is required to keep up with the volume of high quality distributional data being published. Where possible, original data should also be acquired from mammal specialists and integrated. Lemurs and particularly bats are the groups which are presently most data-deficient.
7. *Integrate data in other key groups*.

Table 1. Files included in the zipfiles PDA_GAZ.zip, Butterflies.zip and Mammals.zip

File name	**Size Mb (compr e-ssed)**	**Scope**	**# records**	**Description**
PDA_GAZ.xls	15.6 (4.8)	Madagascar	41,770	Madagascar gazetteer and tools
Mad butt checklist.dbf	0.786 (.050)	Comprehensive	301	Madagascar butterfly taxonomic checklist
PDA butterflies.dbf	44.4 (1.32)	All described butterflies (301 spp./sspp.)	12,062	All validated butterfly data for described species
PDA butterflies header metadata.xls [dbf version also included]	.034 (0.010)	Butterflies	N/A	Metadata* for fields in butterfly database
Mad mamm checklist.dbf	0.470 (.041)	Comprehensive	165	Madagascar mammal taxonomic checklist
PDA mammals.dbf	19.1 (0.61)	Mammals (validated non-volant, volant (bat), expanded taxonomic checklist, and all other input records)	4,837 (2,984 validated non-volant, 285 volant records, 182 checklist, 243 non-georeferenced checklist, and 1223 "excluded" records)	Madagascar mammal data input to date
PDA mammals header metadata.xls [dbf version also included]	.034 (0.010)	Mammals	N/A	Metadata* for fields in butterfly database
Taxonomic checklists metadata.xls [dbf vesrion also included]	0.022 (.005)	Butterflies and mammals	N/A	Metadata for taxonomic checklist field headers

* Note that a distinction between data and metadata (data about data) is made for all fields in these files. This helps in data input because data should not be input into a metadata field.

Reference

Lees, D.C., Kremen, C. and Raharatsimba, H. Classification, diversity, and endemism of the butterflies (Papilionoidea and Hesperioidea) of Madagascar: a revised species checklist. In: The Natural History of Madagascar (S.M. Goodman and J. Benstead, eds.). University of Chicago Press (in press).

Attachment 3: A Brief Report on Botanical Institute Visits by David Lees, March 2001

CONTENTS

INTRODUCTION

Platforme d'Analyse (PDA) is a distributed set of biodiversity information, mapping and prediction tools for Madagascar. Initial efforts in 2000-2001 were made to establish a reasonably comprehensive national database, complete with taxonomic checklists for a couple of moderately well-known faunal groups: mammals and butterflies. In order to have a reasonably comprehensive biodiversity database accessible to those that need it, plants arguably the most important taxonomic groug to incorporate. To this end, discussions with Missouri Botanical Garden (MBG) and Royal Botanical Gardens (RBG) have been initiated. During visits by David Lees in March 2001, existing approaches and mapping/analysis tools were demonstrated. A questionnaire was used to structure feedback, results of which are included here (RBG did not yet formally respond, apparently because their policies on database availability have not yet been finalised). The MBG office at MNHN, Paris was visited on 13 and 16 March 01 (discussions were held with Peter Lowry and George Schatz). RBG, Kew, was visited on 20 March 01 (eight staff came to the meeting). The notes that follow are based on discussions at these informal meetings.

In summary, reaction to the tools and overall concept was generally very positive, whilst some of the botanists' reservations are noted below as regards practical implementation. The most important take-home point is that the botanists would like to play a much more central role in the project than has been the case to date. Mechanisms for integration of the many technical components of the project would need to be worked out in more detail. Two main questions arising were: where the initial database front-end(s) should be housed, and how technically would multiple databases be integrated. Other relevant software developments and approaches elsewhere in the world were discussed and some details and weblinks are provided here for general orientation. MBG, RBG and indeed MNHN (according to their own policies) will make data available anyway, with an emphasis on high quality validated datasets. Data from MBG are likely before RBG, assuming certain minimal conditions can be met.

The two sections of this report focus on the answers to questionnaires received to date from MBG and RBG, respectively. More obvious technical problems are highlighted, and suggested streamlining of the PDA data format to allow greater compatability with botanical data is suggested.

MBG

The public domain data

Names of more or less all described flowering plants of Madagascar are already searchable in the public domain the TROPICOS website: http://mobot.mobot.org/W3T/Search/vast.html.[1][2]. This is about 12,000 plant species in all, of which approximately 1/3 exist with a reasonable taxonomic framework, 1/3, although treated in the flora, are badly in need of revision, and 1/3 remain unrevised. Most of the names (about 26,000) on which this flora is based already exist on the Mobot website, queryable individually. Not only is the data is already online, it is constantly being updated. In terms of actually incorporating data, those taxonomic groups that MBG will contribute to PDA have been checked and validated very carefully, and are thus of the highest quality. Pre-"GPS"-era data for tens of thousands of collections have also been "post-facto georeferenced" (i.e., a decimal latitude and longitude has been added). In the TROPICOS database, this act has been indicated by enclosing the data within square brackets, one possible convention to distinguish interpretations from raw data.

Much other useful work has already been accomplished by MBG and associates in terms of inputting and validating botanical data. For example, notebooks/localities of all the major Madagascar plant collectors have been mined (Humbert alone collected about 23,000 plants). This information has gone to the development of the large (but much more useful than US Board of Names) MBG/TAN/TEF/MNHN gazetteer for all plant collecting localities (see also below, under database structure). This gazetteer is available to the project, and is already public domain on a link that also offers the Madagascar Arcview bioclimate dataset in .e00 format (http://www.mobot.org/mobot/research/madagascar/gazetteer). In fact, PDA could and should play a role in improving this gazetteer by the addition of zoological localities.

The datasets available in the first instance from MBG:
About 600 spp. would be available for release first. The examples given below are examples but by no means a complete list; there are many groups according to Pete that would be good or better.

Araliaceae (75 spp.), about 20% undescribed
Papilionoid legumes (ca. 200 spp.) [With the proviso that I did not manage to meet with Labat this time, but it is important he is asked first].
Endemic families (now 7!) (100 spp.)
A complementary set of species chosen across other groups, with a variety of life forms and other criteria (200-250 spp.)n

[1] Go first to http://www.mobot.org and click on W3Tropicos; make use of TROPICOS options, e.g. "Specimens", to right of screen. Data is not presently available for download *en masse*.

[2] Note that there is also a MNHN database on the web – (SONNERAT http://www.mnhn.fr/base/sonnerat.html)

Minimum conditions for the data available in the first term to PDA

MBG data are available under the following (three) minimum conditions.

MBG would **like a say** in which of the mass of data is downloaded/used first. Specifically, they would require that only data be provided only for groups that meet two main criteria, both of which must be met in order for the data to be of any real value for conservation applications:
a taxonomically solid framework should currently exist; and
collection data must have been compiled and verified from the major herbaria.

MBG would like a **senior partner role** in the project
MBG feel it would be both logical and appropriate for them to play a substantial role in the project, probably focusing on certain components. In particular:

A) they would like to see an **integrated approach**, with the two Madagascan herbaria, TAN [=PBZT] and TEF (FO.FI.FA.-Ambatobe) playing a key role in the project. Globally, 5 institutions (TAN, TEF, MNHN, MBG and RBG) hold 95-98% of all Madagascar specimens – a very fundamental difference from that for zoological specimens facilitated by the availability of isotypes (see also under database structure, below). Obviously, the data is both historical and recent, and covers all parts of Madagascar: this seems to apply to all above-mentioned datasets. (We do not yet have a sense of how many "all-flora" site inventories have been carried out in Madagascar. One would imagine none. Some groups such as palms are underrepresented in collections because they are so awkward to collect; botanists do not rely at all on visual or photographic methods unlike recorders for many biotic groups). 75,000 records are already available in digital format through TROPICOS (many not down to specific level, however- only 40% have been identified to species!). Maybe 40-50% of all specimens remain to be digitised. The five-institution total of specimens is estimated at somewhere under ¼ million. MBG have put on their own numbers onto specimens in addition to the usual field collector's number series (but there is no herbarium accession number); Kew use a unique bar-coding system (see later). It is very important to point out that a lot of this data has already been captured by work by trained Malagasy technicians employed by MBG and based at TAN and TEF, and that their role needs to be acknowledged as crucial. They already have open access computers with links to the TROPICOS database.

B). A second and critical area of importance, as also indicated below, concerns the role of Malagasy institutions holding natural history collections. As pointed out by George Schatz, these collections are THE primary source of biodiversity data for many groups (albeit not the only data for some vertebrate or even invertebrate groups). MBG feel very strongly that any project that seeks to apply biodiversity data to conservation planning must place those collections centrally. It will be necessary at various points to return to the collections to verify and validate certain data, so their role is more than just a source for the initial information that will populate the database. Historically, natural history collections have often drawn the short end of the stick when it comes to support, and the current status of the collections in Madagascar clearly testifies to what has happened locally. Now that people are keen to develop a tool to make data truly useable for conservation, we have what may be a UNIQUE opportunity to showcase these collections, and to show how important they really are. Let's not miss this chance, OK?

3. The third condition is that MBG must be able to play a major role in helping to develop applications of biological data to conservation planning. Not only do MBG have a large and important data component to offer that they are happy to have used, but they also would like to play a part in the process of thinking about how basic biodiversity data can and will actually be analysed and applied to conservation planning. This aspect strongly interests several MBG staff members. Moreover, it is institutionally important that MBG are actively involved in this process. Furthermore, helping to develop applications to conservation planning is a primary motivations. Much of MBG's program development over the last 5+ years has been specifically aimed at building the links for applying verified botanical data to conservation. MBG have a lot vested in what they have done to date, and strategically it is critical that they continue working in this direction

Funding requirements and budgetary constraints

MBG have NO direct funding requirements from the project, except that they would like to see a collaborative approach. If PDA is going to be making its own proposal, we should say that the planned botanical proposal by MBG to CEPF would facilitate collaboration on PDA (coordinated and complementary objectives should be demonstrated). MBG will also submit their own proposal somewhere for personnel to complete data entry to the minimum standards of PDA. There is clearly the need to compare notes so that any MBG CEPF proposal (they would go for a three-year project) could be complementary to any application made on behalf of the PDA.

MBG DATABASE STRUCTURE

See attached suggestions for modification of the PDA format.

Notes from Peter Lowry (in lower case):
CLASS AND ORDER: of little use for angiosperms
SP_AUTHOR: for plants, the author of the basionym and the combination are needed [In TROPICOS these are separate fields, linked to a comprehensive table that contains author names and the accepted abbreviations]
INFRASPECIFIC: variety is needed too
SYN_AS: what to do about multiple synonyms?
SP_UNCE: botanists don't use this at all.
TYPE: add isotype (I), lectotype (l) neotype (n); a specimen can be a type of a name that has been placed in synonymy under the name currently accepted. How will such case be handled?
ORIGINAL NAME: needed here if a type!
COLLCTR: must be indexed very carefully! [Same may apply to locality! For example, the surname with some convention for initials and extra collectors, or the most commonly used minimal locality unit – e.g. "Andranobe", not "Madagascar Nord-Est" - should come first]
CLCT_MTH: n/a for plants
CLCT_DAT1, CLCT_DAT2, LCTY_NAME, SITE: – link to MBG gazetteer here?
SITE_CODE: not used for plants
LAT_DD: between LAT_SEC and LON_DEG [?]
ELEV_MAX: min also?
HAB_DESC: ultimately should be standardised for new data…. MBG to work on this
MUS_CODE: (when available)
PLACE OF PUB OF ACCEPTED NAME: Most recent taxonomic revision?
IMAGE: field for images (photos, digitized images), some of which available on web (http://www.mobot.org/MOBOT/Madagasc/welcome.html).

Comments and reservations

This section stems largely from the discussion on 15th March.

Peter Lowry is very positive about this project. Indeed, from earlier communications, he sees it as perhaps a unique chance for Madagascar to get its act together to apply for funding for biodiversity databasing, fitting in very well with global initiatives too, like GBIF (Global Biodiversity Information Facility). It would position Madagascan institutions to get maximum benefit from CBD (Convention of Biological Diversity), now it has endorsed GTI (Global Taxonomic Initiative). He sees that we can move forward in a complementary fashion. In his view, MBG will only lose out if they do not join the political process that is already in train with PDA. Pete emphasizes how important it is to set the database standards high to start with (indeed, without properly validated data, one cannot possibly expect to be able to make valid conservation planning decisions).

George Schatz takes a longer-term, more skeptical view. The tendency of Malagasy institutions to want to "possess" a project of this nature is a potential stumbling block. Whilst adamant that the database should not be based at a single office or institution, he suggests that the major herbaria should play a key role, where for example PDA could draw directly from TROPICOS.

George points out that other software is available. There is Species Analyst at Kansas Natural History Museum in association with the San Diego Supercomputing Centre (see http://habanero.nhm.ukans.edu/TSA; http://www.npaci.edu/online/v3.19/species_analyst.html and other links such as to "GARP" and its latest incarnation WhyWhere). (According to Claire, this is good for pulling records from diverse sources, rather than the basic mapping side. DL investigated this in Sept-2001 and there seems to be an unexplained bug with the SppAnalyst.ini file which prevents the extension working properly with Arcview3 at least on Windows98). The Danish working in Ecuador have done a lot of nice work over the past 5 years. See Skov, F. 2000. Potential plant distribution mapping based on climatic similarity. *Taxon* **49**: 503-515, and references therein. The recently published Libro Rojo of the endemic plants of Ecuador (Valencia, R., N. Pitman, S. León-Yánez & P. M. Jørgensen (eds). 2000. Libro rojo de las plantas endémicas del Ecuador 2000 [Red Book of Ecuador's Endemic Plants], PUCE editions, Quito.) should serve as model for countries everywhere. Nigel Pitman has done a magnificent job assisting the Ecuadorans. As is the case in most areas, the Australians are way ahead of everyone else. Mike Austin's and Hutchinson's work (ANUDEM, ANUSPLIN, ANUCLIM), as well as a new conservation planning tool called C-Plan (an interactive, iterative program that incorporates complementarity and irreplaceabilty measures). And finally there is DOMAIN (Carpenter *et al.*, 1993, Biodiversity and Conservation **2**: 667-680), a flexible modelling procedure for mapping potential distributions of plants and animals.

These discussions suggest that whilst the PDA software is actually already customised to a specific task (range mapping and polygon query based on point data), it will need to add gap analysis functionality and make use of different levels of record reliability as priorities in near-term development plans. More importantly, it needs to go web-based.

Also taxonomy needs to become much more prominent in conservation planning in Madagascar, and indeed the taxonomic institutes such as herbaria much more prominent in PDA. Indeed, Peter Lowry sees this project as a real chance to "sink the hook" in terms of getting taxonomy regarded as of fundamental importance..

George would like to go much further, and see founded an association of natural history collections in Madagascar. A single autonomous Antananarivo natural history institute would combine all the major biotic interests. For example, a policy forum paper by Stuart Pimm *et al.* for Science estimates the costs of conserving biodiversity. One of the major points of this paper concerns the establishment of an additional 25 biodiversity centres such as INBio, CONABIO, and Humboldt Institute in the "hotspot" regions and the most critical countries. From the draft: "..Budgets for well-established and effective centers are a few million dollars per year. Roughly half a billion dollars would support 25 centers for a decade .." Madagascar institutes need to move forward and be helped to establish such a centre. It is absolutely critical that any sort of Biodiversity/Conservation Planning database/tools be fully integrated into local taxonomic initiatives and natural history collections, NOT just local governmental and non-governmental environmental/protected areas management/federal lands management organizations, i.e., ONE, ANGAP, MPAEF and foreign conservation NGO's (CI, WCS, WWF). Long-term success depends upon local taxonomic capacity. What Madagascar really needs in his view is a National Museum of Natural History (or some such institution) where ALL collections can be housed, properly cared for, and expanded through integrated inventory work. They can also be used for a whole range of applications, including research and various applications – including the ones we are considering here. This is much broader than just taxonomy. It involves all the facets relating to collections, of which taxonomy is just one.

****

RBG

Staff present at meeting
Petra Hoffman. Madagascar Geographical Officer
Justin Moat. Head, GIS unit
Aaron Davies. Herbarium, Rubiaceae.
Paul Wilkins. Herbarium. Dioscoreales.
Madeline Harley. Palynology.
Wolfgang Stuppy. Living Plants Collection.
John Dransfield. Herbarium. Palms

The public domain data

I don't think there is any yet, although Kew's own website has some GIS related themes for download (http://www.rbgkew.org.uk/herbarium/madagascar/).

The datasets available in the first instance from RBG:
Palms (ca. 170 spp.), 1500 records [ALREADY PUBLISHED]

Legumes (Papilionoideae: 422 spp. + introduced taxa; Caesalpinoideae + Mimusoideae: unknown # spp.), 10000 records [IN PRESS]

The datasets available later from RBG:
Rubiaceae (?? spp.) 6000-9000
Dioscoreales or "yams" (<50 spp.?) 1000 records

How the datasets are going to be available:
They cannot be released without the permission of Mark Jackson, Kew Databasing Manager (responsible for data standards and integration, and head of the Scientific Databasing Committee) who referred my request to Alison Prior (A.Prior@rbgkew.org.uk). It has not yet been possible arrange a meeting with either person, but the provisional answer is that Kew has not yet made a global decision about the intellectual property issues of data contribution, and it is not clear at this stage when such a decision will be made.

The existing database example is SEPASAL (a drylands project: http://www.rbgkew.org.uk/ceb/sepasal/). This is a good example for connectivity across the web. It has user logins for variable levels of access. Names are put in another database IPNI (International Plant Names Index: http://www.ipni.org/) which has user control and covers wide areas- e.g. Australia.

It was mentioned at the meeting that Kew has a proposal in the offing to UK government for development for a package potentially called EPIC (Electronic Plant Information Centre). This will be software that integrates databases and can be used as a search engine, as well as pushing data onto the web. It would be potentially very relevant for what we are planning to do, if it was available in time. All Kew's data contributions would, in theory, go through EPIC.

Restrictions and other constraints to contribution
The availability of an up to date vegetation map is a major problem for reliable range mapping. Kew hope to get funding for doing this. MBG also plan to get some simple standardisation in place for the description of new information about vegetation for new collections: there is nothing standardised at present.

Endangered species – particularly succulents, palms [19 extinct/endangered, 50 endangered, 39 vulnerable] and orchids.

The need to go one step beyond the raw data: Kew need to be able to put their own filtering and interpretation on the data to make sure it is not misapplied or misused.

Field for restricted data [from internal side] – as required by CBD or own institution's intellectual property rights. May be text about restriction-agreements, permit nos., etc.

Minimum conditions for the data available in the first term to PDA

See above.

RBG DATABASE STRUCTURE

Fields critical for taxonomic group in question

HISPID ("Australian herbarium") standard – see http://www.rbgsyd.gov.au/HISCOM/HISPID/HISPID3/hispidright.html

CORE FIELDS approach desirable
This approach may have some benefit for simplifying our database structure and for facilitating data entry, so that users can just add on the more specialised field. For example, if one is keying in museum specimens of butterflies, one only needs the following (around 20) fields: Genus, Species, One or more infraspecific fields, Sex (if one specimen per record), Voucher info on specimen [e.g. dissection number], Identity label on specimen (an important omission!), Type, Original name of type [this is an important omission!], Collector name, Collection date(s) [day month year in separate field, if available], Locality name including elevation [both of these two should be parsed on the trot to optimise indexing) [I feel strongly that elevation if available is part of the entire locality description], Latitude and longitude if given, Comments on specimen or taxon, Museum or Herbarium Museum name, Museum accession code [more often used for bequests of zoological specimens; often includes a number and year, e.g. 78-20 which might be the twentieth accession in the register in 1878; it is not guaranteed to be unique but may be read off the data label in many cases] , Museum or herbarium accession date [this is important for historical data- may be the only date available! May need to be an interpretation of the accession code], Unique specimen number [this is an important omission! – there is little point databasing a specimen if it cannot be tracked]. If one is entering literature data, one wants a much wider range of field that may include datasource fields.

Note: Claire has pointed out that there seems to be a lot of confusion about different number series. It is vital that the number is really unique, and in practice this is difficult to achieve (e.g. a collector's individual number cannot be guaranteed to be unique), and may therefore need a number/code to be applied afterwards. Some institutes (e.g. NHM, RBG) already have a system (that I have to say is not yet always flawless, e.g. duplicates in some systems can be accidentally issued); others (e.g. MNHN Entomology) do not appear to have yet even thought through the issues.

Here's Kew's core fields- please note though, this is designed for internal staff requirements (e.g. Country, Kew Geographical Region are not needed in PDA).

Specimen data

Unique number [barcode]
Family
Genus
Species
Infraspecific epithet
Country
Kew Geographical Region
Primary collector's name
Collector's identifier/number
Restriction
Collection date (in three fields)
Locality information [exactly as on label]

GIS

Coordinates as below: note that Kew made a decision that because of the default of many GPS systems to decimal minutes [not degrees], it was essential to force people to enter degrees minutes and seconds and orientation (the latter rather than forcing positive or negative at the time of entry. People's views on this? [I would say, as we are focusing on Madagascar only, we don't worry about orientation].
Lat_deg
Lat_min
Lat_sec
Lat_or (N or S)
Lon_deg
Lon_min
Lon_sec
Lon_or (E or W)
Source of lat/lon [Lookup: Label (default); Map estimate; GPS; Gazetteer; Internal Gazetteer; Literature; Other]

What to do with morphospecies? In the case of Rubiaceae, Aaron Davies (see http://www.rbgkew.org.uk/herbarium/madagascar/mad_rub.html) points out these can constitute maybe 40% or more of data, and so he felt they should be included for conservation planning. Nevertheless, the consensus in the group was that they have to be described first.

Acknowledgements

Kew would like their logo used, and also make acknowledgement database dependent. There should also be a list of contributors.

Additional tools

Control on how locality is used
Stats on records, etc. (Justin Moat implements Chi-squared tests)
RBG has developed their own customised GIS software for taxonomic data query, range mapping and threat evaluation and have incorporated dynamic links with MS Access, which comes complete with its own data entry facilitating modules.

Other existing software was mentioned. Flora map produces predicted maps based on Map Objects (http://www.esri.com/software/mapobjectslt/description.html).

Funding requirements and budgetary constraints

No funding was specifically requested from PDA by RBG.
However for CEPF, a complementary rather than competitive strategy was suggested.
Petra Hoffman mentioned that Kew wanted to get some funding to capture specimen data in Paris, but hadn't identified a source for this yet. PDA could possibly help here, one wonders if the target group(s) were identified?

Justin Moat was asked theRBG GIS department would potentially be interested in a role furthering development of the PDA. This would depend on external funding.

Comments and reservations

See above.
Also, Justin Moat's major query with PDA is is exactly how efficient integration/pooling of multiple databases would be achieved. A metadata clearing house for data is one thing (i.e. showing who has what); but integration could be more complex. Much will depend on the goal of PDA. Also can anyone predict at this stage how much work is going to be involved and how much it will cost?. (Note that possible solutions come out of Claire's AMNH trip. A software engineer will be recruited to deal with the technical integrative and development problems. Each dataset would be pre-validated by the

donor/museum/herbarium, geo-referencing in particular solved by their own taxonomic experts who understand limitations of gazeteers and locality names. The validated dataset would have one code indicating whether it can be released to PDA, another indicating the recommended access level for others. Also, each data donor would come up with its one field mapping to PDA for automated update. It would remain for the steering committee to decide what constitutes valid data). Examples of other potential technical problems that a recruited software technician would need to solve: deciding on core fields, incorporating live links and data updates, filtering problems, provisions for long-term database maintenance by each data donor.

Consider a specific problem raised by Justin for updating localities. Many localities (e.g. villages) relocate gradually over time. Thus you can cause more problems than you solve if you automatically update a collector's locality, when the original co-ordinates might actually be different. That is the reason they decided not to auto-update locality geo-references at Kew. (To me, this raises the question as to whether a database should have more than one set of coordinate fields). Kew have one nice innovation with regard to the gazetteer which tackles the problem of multiple homonymy of localities, particularly severe in Madagascar. They have a hit mechanism which shows how often a particular locality homonym (let's say, Ranomafana) has been selected. Otherwise, using lookups for gazetteers in Madagascar is fraught with difficulties. In fact, putting in latitudes and longitudes through a gazetteer could be argued to be a pointless exercise unless done by experts on both geography and collecting localities. (In my own use of a gazetteer, I simply put in bold the most regularly used one). The update and maintenance of a web gazetteer obviously needs some more consideration, then.

Appendix 1. PDA Minimum Standards.

We hope that institutions will be able to fill out most of the fields for most records, as appropriate. If unknown, leave field blank. Optional fields are included in further tables below, depending on purposes.

Below follows modifications to our previously suggested structure, based largely on the information from MBG/RBG.

Suggested PDA core fields for terrestrial animals and plants:

FIELD NAME	Explanation
FAMILY	Taxonomic family [facilitated by lookup]
GENUS	Taxonomic genus [facilitated by lookup]
SPECIES	Taxonomic species (note, if unknown, do not put species or question mark here)
SUBSP	Taxonomic subspecies [mainly for animals]
INFRASPEC	Other infraspecific entities, e.g. plant variety
SP_UNCE	Taxonomic species status questioned in source (e.g. _?, _cf., _aff., sp._nov., sp., sp.A, morphospecies)
TYPE	Type status if known = holotype (HT); paratype (PT) [= "allotype"]; syntype (ST) ; lectotype (LT); paralectotype (PLT) and for plants: isotype (IT); neotype (NT). (If blank, does not pertain to a type). The prefix "ms" = manuscript, in the case that type status does not appear to have been published. Designated by whom, if known (usually it will not be known).
TYPE_NAME	Original name of type. In the case of multiple types (a single specimen can be the type of more than one name) it is strongly recommended that the type name (and sex, at least for some groups) is added, e.g. in square brackets into the TYPE field. In fact, it might be better if both fields were combined in a single text field
VOUCHER	Text and code for voucher (e.g. morphological preparation, DNA etc., photograph, etc.)
NO	Number of specimens lumped under record [for animals]. With museum/herbarium specimens, we would go to specimen level if possible, for taxonomic reasons alone.
SEX	Sex if known, if a single record [mainly for animals, especially important for insects]
COLLCTR	Name(s) of collector(s) or recorder(s) of specimen [The surname should precede the initials or christian name, but leaving no blanks, e.g. Grandidier_A]. See also remark under MUS_COLL
COLLCTR_NO	Original collector accession number/code if available. This will usually be a collector's field number, not be confused with the

FIELD NAME	Explanation
	museum accession code (MUS_CODE) and the unique specimen tracking number (UNIQUE_NO)
CLCT_DAT	Collection date [must be split into 3 fields, corresponding to day, month, year; also desirable to have 2 additional fields, where sampling covers period: CLCT_DAT1 and CLCT_DAT2; CLCT_DAT1 is automatic IF 1ST 3 FIELDS FILLED
LCTY_NAME	Locality name [Data entry person ideally needs to be trained to parse information before entry, with minimal recognisable name corresponding to the locality first, followed by successively larger geographic units delimited by commas, followed by elevation, if available. Otherwise, unparsed label information goes in label_info.]
LAT_DEG	Latitude degrees
LAT_MIN	Latitude minutes
LAT_SEC	Latitude seconds
LON_DEG	Longitude degrees
LON_MIN	Longitude minutes
LON_SEC	Longitude seconds
COO_SRC	GPS/MAP/GIS/GAZ/OTHER. If gazetteer specified, gazetteer from which derived
MUSEUM	Code name of museum or herbarium (e.g. AMNH, PBZT, MNHN, BMNH, USNM, TAN, TEF, MBG, RBG)
MUS_COLLN	Name of collection or bequest (multiple entries separated with forward slash with earliest first, if there is a collection history, as there often will be with insects. In many cases it may not be clear at the time of data entry if the original collector/collection name is the same as the field collector)
MUS_CODE	Museum accession code number (this will be the historical accession number, often associated with a year; not guaranteed to be unique: eg. 78-25)
MUS_DATE	Museum accession date, in square brackets if inferred e.g. from MUS_CODE (such as [1878] 78-25)
UNIQUE_NO	Individual specimen unique tracking number (ideally part of block issued by institute housing collection e.g. museum registration number, barcode). In many cases a unique number will need to be assigned at the time of data entry, provided it can also be placed on the specimen. This is not to be confused with the Museum accession code number.
VALIDATED	Name of person validating the record and date validated, if validated
RESTRICTION	Identifies restriction placed on record by donor (CBD requirement)

FIELD NAME	Explanation
LABEL_INFO	Label information (exactly as on label)

Accessory fields required for particular purposes – for example, if the data input is from literature, the data entry fields below will be needed or called up

Literature and web datasources

TYPE_PUB	J= journal article, B= book, C= chapter in book, R= report, T= Thesis, E =electronic
AUTHOR(S)	(essential to format consistently as for COLLCTR(S): Goodman_SM; Carleton_MD; etc)
ART_TITLE	Title of article or book chapter
EDITOR(S)	Editor(s) of book for book chapter (format as above)
PUBL_DATE	Publication date (year)
BOOK_TITLE	Title of book
PUBL	Publisher of book/Institute for report
CITY_PUBL	City of publication
JOURNAL	Journal title
VOLUME	Volume number
PART	Part, issue, series, tome
PAGES	Pages (in format « 233-267 » or « 245 pp. »)
CITE_AUTH	Authors of specific work cited within the main reference (format as above)
CITE_DATE	Year of specific work cited within the main reference
URL	Url of an electronic publication
IMAGE	Link or url of image (e.g. on web) or code for photo

Frequently used accessory observational fields:

COMM_TAX	Comments on specimen such as distinguishing/diagnostic features observed on specimen
TAX_PUB	Literature source for accepted name (e.g. most recent taxonomic revision)
IDENTIFN	Identity or determination labels applied to specimen [can be important for types]
CLCT_METH	Collecting method [for animals]
SAMP_TYPE	Sample type: POINT, PLOT, TRANSECT, TRAIL, AREA
SAMP_DIST	Half-length of transect or trail; or radius of plot or area in Samp_Type (km.)
SAMP_SHP	Name of shapefile associated with Samp_Type, if any
ELEV	Elevation, if recorded at location (m.)
ELEV_MAX	The max elevation of record, if available (m.)(Elev_Max = Elev if it is a single point)
ELEV_SRC	How elevation was recorded (GPDC/GPSUC/ALTIM/MAP/GIS/GAZ/LIT {LIT=default from source

	info, whether literature or not}). Note that GIS calculation is covered by Elev_calc
HAB_DESC	Description of habitat, particularly vegetation type at location. This is not intended to be standardized, but simply comes from field notes.
HAB_CD	Code indicating vegetation type at location: Suggested codes: FOREST HABITATS: D= Deciduous forest; Dl= Western littoral forest; Du= deciduous *Uapaca bojeri* (Tapia) formation; R= Rainforest; Re = Ericaceous (*Philippia*) forest; Rg= Gallery (riparian) forest in otherwise dry habitat; Rm= "montane", Rl= Eastern littoral forest; Rs= Swamp (seasonally inundated) forest; Rt= Humid forest transitional between rainforest and deciduous forest ; T= Thorn scrub/ thicket (Didieraceous or euphorbiaceous formation); M= Mangrove forest. OPEN HABITATS: Sn= semi-natural grassland/meadow; Sd= Degraded savannah; Sw= wooded savannah; X= Cleared forest (any type); N= Town or village; E= Roads/Roadsides; A= Sand dunes; B=Cliff/crags/rocky outcrops; C= Beaches; F= Mud-flats. AQUATIC HABITATS: Qf= River/stream with forest; Ql= Lakes; Qm= Marshland; Qr= River/stream without forest; Qs= seasonal standing water (Matsabory). SUBSURFACE HABITATS: H= Soil; I= Leaf litter; V= Cave. AGRICULTURAL HABITATS: Ow= Wet crops e.g. paddy ricefiel
DIST_DESC	Description indicating habitat disturbance type at location. This is not intended to be standardized [although ultimately it may be], but simply comes from field notes.
DIST_CD	Code indicating habitat disturbance type at location. P= pristine or semi-pristine; S= regrowth forest beyond stage of savoka; D= heavy disturbance: Ds= "savoka"/jachere; Dt= "tavy" - recently cut forest/clearing; Dp= Domestic animal pasture "Fahatr'Omy";

Most of the groups of fields below can be assigned to the database *a posteriori* and so are not minimum requirements

Taxonomic and threat status checklist

CLASS*	Taxonomic class [for animals]
ORDER*	Taxonomic order [for animals]
FULL_NAME*	Genus and species name only for indexing
SP_AUTHOR*	Including author of combination (for plants)
SYN_AS	Taxonomic synonym (put original name here if it is known to be a synonym)
ENDEMIC*	Indicator if species is endemic to island of Madagascar E= Endemic Madagascar MR= Endemic Malagasy Region OW= Old World AT = Afrotropical PT= Pantropical

IUCN_CODE*	IUCN status code:EX=extinct,EW=extinct in the wild,CR=critical,EN=endangered,VU=vulnerable,LR=low risk (3 subcategories within LR: cd=conservation dependent; nt=near threatened; lc=least concern, removed from 1996 Red List) DD=data deficient,NE=not evaluated. Here using IUCN 1996 Red List of Threatened Mammals
CITES_CODE*	CITES status code: App1,2 etc

Accessory geographic information

SITE	Site name within given locality
SITE_CODE	Code of site within locality
LCTY_PA^	Name of protected area, if applicable
LCTY_PROV^	Province (FIAN= Fianarantsoa, ANTS= Antsiranana, ANTA= Antananarivo, MAHA= Mahajanga, TOLI= Toliara, TOAM= Toamasina)
LCTY_FIV^	Fivondranana
LON_DD*	Longitude decimal degree
LAT_DD*	Latitude decimal degree

Completely redundant field

MUS_LOCN	Location of museum or herbarium (city/country)

NOTES

In general. This format is intended to provide a minimum standard, but remains flexible for the time being. We are open to suggestions on adding fields or altering explanations, particularly for specialized needs related to specific taxa.

1. Any interpretation must be enclosed in square brackets

2. (^ signifies GIS will calculate by intersection with lat and lon; * signifies database module will automate)

3. All datafields should be strings and not restricted in length (e.g. Access can take up to 255 width without wasted storage space).

4. Also note: the decimal lat and lon will be calculated automatically and mapped to the output database file for PDA to use, so there is no need for the inputter to add this information, except in the case of GPS data. Great care is needed for GPS output that stores decimal minutes and such problems can be dealt with in a databasing module. Most historical data will lack primary or reliable observations of longitude and latitude anyway, so it is not usually necessary for the data provider to provide this information, although we can

take decimal longitude and latitude as an alternative to degrees, minutes, and seconds.